AgRISTARS

IT-T3-04395

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

JANUARY 1983

Inventory Technology Development

FINAL REPORT: AREA ESTIMATION USING MULTIYEAR DESIGNS AND PARTIAL CROP IDENTIFICATION

R. L. Sielken, Jr.
Texas A&M University
Institute of Statistics
College Station, Texas 77843
NAS9-13894

(E83-10195) AREA ESTIMATION USING MULTIYEAR DESIGNS AND PARTIAL CROP IDENTIFICATION Final Report, 1 Mar. 1981 - 31 Dec. 1982 (Texas A&M Univ.) 16 p HC A02/MF A01 CSCI 02C G3/43

N83-20317

Unclas 00195

Earth Resources Applications Division
Lyndon B. Johnson Space Center
Houston, Texas 77058

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
IT-T3-04395		

4. Title and Subtitle	5. Report Date
FINAL REPORT: Area Estimation Using Multiyear Designs and Partial Crop Identification	January 1983
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No
R. L. Sielken, Jr.	
	10. Work Unit No.

9. Performing Organization Name and Address	11. Contract or Grant No.
Texas A&M unversity Institute of Statistics College Station, Texas 77843	NAS9-13894
	13. Type of Report and Period Covered

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
National Aeronautics and Space Administration Earth Resources Applications Division Lyndon B. Johnson Space Center Houston, Texas 77058 Tech Monitor: M. C. Trichel	March 1, 1981 to Dec. 31, 1982
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This final report refers to project number 3054 entitled "Area Estimation Using Multiyear Designs and Partial Crop Identification". This project spanned the period from March 1, 1981, to December 31, 1982, and is the last project undertaken under contract number NAS9-13894. The earlier projects under this contract have been reported in previous final reports.

During the project period work has been focused on the following three areas:

1) estimating the stratum's crop acreage proportion using the multiyear area estimation model,

2) assessment of multiyear sampling designs, and

3) development of statistical methodology for incorporating partially identified sample segments into crop area estimation.

17. Key Words (Suggested by Author(s))	18. Distribution Statement

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified		

*For sale by the National Technical Information Service, Springfield, Virginia 22161

During the project period work has been focused on the following three areas:

i. estimating the stratum's crop acreage proportion using the multiyear area estimation model,

ii. assessment of multiyear sampling designs, and

iii. development of statistical methodology for incorporating partially identified sample segments into crop area estimation.

Although each of these areas is reviewed separately below, the overall goal of improved crop area estimation utilizes all three areas jointly.

Our objectives in this project have been more than met. We have developed and documented the statistical methodology needed to utilize the multiyear area estimation model to produce a good estimate of the stratum's crop area proportion based upon current and previous years' estimated crop area proportions in sample segments. By assessing the impact on the stratum's crop area estimate, we have derived recommendations for how the sample of segments should vary from year to year . Finally, we have determined and tested procedures for explicitly utilizing only partially identified acreages as well as sample segments with completely identified crop acreages.

The three aspects of our research under this project have been separately documented in our technical reports 20, 21, and 22. Dr. R. L. Sielken Jr. gave two invited presentations on our research at the Joint Statistical Meetings of the American Statistical Association and The Biometric Society in Cincinnati, Ohio, during August 16-19, 1982. These presentations were entitled "Multiyear, Through-the-Season Crop Acreage Estimation Using Estimated Acreage in Sample Segments" and "Incorporating Partially Classified Sample Segments Into NASA Acreage Estimation Procedures". Dr. E. E. Gbur also gave a presentation at the same national meetings entitled "Rotation Sampling Designs" which reported on our

research concerning multiyear sampling. All three of these presentations are being published in the Proceedings of the Section on Survey Research Methods.

1. Estimating the Stratum's Crop Acreage Proportion Using the Multiyear Area Estimation Model

The basic model relating the stratum's at harvest crop acreage to the crop's estimated at harvest acreage in the sample segments has the general form

$$y(\text{observation}) = \text{year effect} + \text{segment effect} + \text{season bias} + \text{noise} \qquad (1)$$

where $y(\cdot)$ is an appropriate transformation. The specific form of model (1) is

$$y(p_{ts\ell}) = \alpha_t + b_s + \ell_\ell + e_{ts\ell} \qquad \begin{matrix} t = 1,\ldots, T, \\ s = 1,\ldots, S, \\ \ell = 1,\ldots, L \end{matrix} \qquad (2)$$

where

$p_{ts\ell}$ = the estimated proportion of the s-th segment's acreage that will contain the crop at harvest time in the t-th year when the estimate is made at crop calendar time ℓ (for example, ℓ=1 could denote early season, ℓ=2 mid-season, and ℓ=3 at harvest time);

$y(p_{ts\ell})$ = a variate transformation of $p_{ts\ell}$;

α_t = the stratum's transformed crop acreage proportion for the t-th year;

b_s = the s-th sampled segment's departure from the stratum's transformed crop acreage proportion; the b_s's are random variables with expectations zero and variance σ_b^2;

δ_ℓ = the systematic difference between the non-harvest time estimates of the crop's transformed at harvest acreage proportion and the corresponding estimate made at harvest time ($\delta_L \equiv 0$);

$e_{ts\ell}$ = the aggregate of sampling and classifications errors in the transformed data.

The primary objective is to estimate the crop's at harvest proportion of the stratum acreage in the current year, T; that is, estimate the inverse transformation of α_T denoted by $P_T = y^{-1}(\alpha_T)$. Secondary objectives could be improved estimates of at harvest acreages in previous years or estimates of changes in the stratum's crop at harvest acreage proportion from year to year.

Estimates of the stratum's crop at harvest acreage proportion are often needed throughout the current year as well as at harvest time. For example, an early season estimate based on observations for ℓ=1,..., L for t=1,..., T-1 and only ℓ=1 for t=T is often desired.

Of course, even though the estimate $\hat{P}_T = y^{-1}(\hat{\alpha}_T)$, of the stratum's crop at harvest acreage proportion for the current year involves only $\hat{\alpha}_T$, the $\hat{\alpha}_T$ depends on the entire multiyear data set and the estimates of the segment effects and the systematic biases which are assumed to be constant from year to year.

The simplest transformation, y(p), of the estimated segment crop acreage proportion, p, to use in (2) is the identity transformation

$$y(p) = p \ .$$

However, it is very doubtful that the additive model (2) would hold for y(p)=p particularly if the p's exhibit a large variation within the stratum. On-the-other-hand a multiplicative model for p may be more reasonable and a logarithmic transformation, $y(p) = \ell n(p)$, more appropriate. The logit

transformation,

$$y(p) = (1/2)\,\ell n[p/(1-p)],$$

is another useful transformation which approximately converts a multiplicative model for p into an additive model for y(p). All three of the above transformations are considered in Technical Report No. 20. There approximate expressions are derived for

(i) the variance of y(p),

(ii) the bias of $\hat{P}_T = y^{-1}(\hat{\alpha}_T)$,

(iii) the mean squared error of $\hat{P}_T = y^{-1}(\hat{\alpha}_T)$, and

(iv) confidence intervals for P_T

under the assumption that p arises from a binomial random variable.

When estimating the parameters $(\alpha_t, b_s, \delta_\ell)$ in model (2), it is not particularly reasonable to assume that the variance of $y(p_{ts\ell})$ is the same for all t,s,ℓ. Hence a weighted least squares analysis procedure has been derived as opposed to the usual unweighted least squares procedure.

A self-contained computer implementation of the weighted least squares estimation procedure has been given to Lockeed, NASA, and ERIM.

Research is continuing under a new contract on several related issues. The sensitivity of the estimate, $\hat{P}_T = y^{-1}(\hat{\alpha}_T)$, of the stratum's at harvest crop acreage proportion to such things as the transformation used, the accuracy of the weights, and the reliability of the estimate of $\gamma = \sigma_b^2/\sigma_\varepsilon^2$ is under study. The empirical behavior of the approximate expressions for the bias of $\hat{P}_T$ and the mean squared error of $\hat{P}_T$ as well as the approximate confidence intervals on P_T is also being evaluated. The extension of the basic model (1) to include year-segment interactions and segment-season interactions is being considered. Another possibility is to replace the seasonal bias term in (1) by a covariate in terms of something like the

number of "crop calendar days" passed by the date of the last satellite imagery used in determining the estimated segment at harvest crop acreage proportion.

Another important line of research concerns the nature of the weights themselves. If the true segment at harvest crop acreage proportion were p^* and the estimated p's were binomial in nature, then the variance of a segment estimate p would be proportional to $p^*(1-p^*)$. Furthermore, the variance of $y(p)$ could be derived for a given y, and the appropriate weight in the weighted least squares procedure could be straight-forwardly approximated using the estimated p. However, the variance of the estimate of the segment's at harvest crop acreage proportion may not be binomial in nature but rather depend on such things as

(i) the satellite being used,

(ii) the sharpness of the satellite imagery,

(iii) the amount of satellite imagery available at the time of the segment estimate.

(iv) the nearness of the segment's observed behavior to classical crop profiles,

(v) the season during which the estimate is being made,

(vi) the weather conditions during the crop's growing season,

(vii) the composition of the segment, etc.

The derivation of appropriate weights under this latter scenario is being investigated.

2. Assessment of Multiyear Sampling Designs

In general, we have a population of segments which is to be sampled for T consecutive years. In any proposed sampling design, the units to be sampled can change from year to year but not at time points within the year. In addition, there is a positive correlation between the responses from a segment in consecutive years which can be utilized to reduce the standard errors of the estimators of the end of year means. The problem is to determine a T year sampling scheme which is optimal in some sense.

In assessing possible multiyear sampling designs we assumed that the eventual estimation would be based upon the multiyear model (2) discussed in the preceding section. Our conclusions were derived from analytical results for particular situations and from exhaustive enumeration of all possible sampling designs for $T = 2,3$, $L = 2,3$, $R = 2,3,4,5$, and $\gamma = .25, .5, 1.0, 2.0, 4.0$ (where γ is the ratio of the variation between segments to the variation between observations on the same segment due to measurement error). Extensive simulations were also performed to determine the distributional characteristics of the estimators under different sampling designs.

Technical Reports 18, 19, and 22 describe for two and three year sampling designs the behavior of the estimator of the stratum's at-harvest crop acreage proportion in the last year of the design. Technical Report 18 obtains a numerical efficiency for each two or three year sampling design for the case where all segment observations have the same variance and hence the weighted least squares estimator becomes simply a least squares estimator. Technical Report 19 generalizes these results by considering the case where the variances of the observations are not necessarily all equal. Here the more efficient sampling designs from Technical Report 18 were compared in terms of the

distributions of the corresponding stratum crop acreage proportion estimators. Finally, in Technical Report 22 these more efficient sampling designs were compared in terms of the distributions of the corresponding stratum crop acreage proportion estimators when cloud cover, etc. caused a random occurrence of missing segment observations. This last study most closely reflects reality. Specific sampling design recommendations are made in the individual technical reports and are not recounted here.

In the paper "Rotation Sampling Designs" Gbur and Sielken discuss two optimality criteria for sampling designs which depart from the criterion considered in Technical Reports 18, 19, and 22. One of these criteria reflects the desire to minimize the average variance of the at-harvest crop acreage proportion estimator where the average is taken over all years instead of just the last year. The second criterion reflects the desire to minimize the variance of linear combinations of at-harvest crop acreage proportion estimators over more than one year - for example, a desire to minimize the variance of the estimated change in the stratum's at-harvest crop acreage proportion from one year to the next. These two criteria do not necessarily lead to the same "optimal" designs nor do they always lead to the same "optimal" designs discussed in Technical Reports 18, 19, and 22.

During our assessments of sampling designs it has been observed that for almost any good two year design there is an extension of that design to a third year which is at least a near-optimal three year design. This naturally suggests a sequential approach to constructing the sampling design. In our new contract we will undertake to develop and implement computer software capable of sequentially constructing next year's design utilizing the specified sample size for that year (possibly different from preceding years)

and utilizing the crop acreage proportion information gathered thusfar as well as the information on which sample segment observations were missing.

3. Utilizing Partially Identified Sample Segments

A small sample of segments within a large region is selected. Each sample segment is observed via satellite at several different times during the crop growing seasons. The objective is to estimate for each crop of interest the proportion of the region's acreage corresponding to that crop's harvested acreage.

In Technical Report 21 we assume that there are only two crops of interest. Furthermore, we assume that only data from the current growing year are to be used in estimating the crop at harvest proportions. The cases where more than two crops are of interest and/or data is available from more than one growing year will be considered in future research.

The sample segments are all assumed to be of the same size. No assumption is made about the region size or the segment size. The sampled segments are assumed to represent a random sample (without replacement) from the segments in the region.

Each sample segment is assumed to have been observed at least once during the growing year and possibly several times. The two crops of interest are designated as crop A and crop B. When a sample segment is observed, the observation can have the form (p_A, p_B, p_{other}) where

p_A = the estimated proportion of the segment which will be harvested in crop A,

p_B = the estimated proportion of the segment which will be harvested in crop B, and

$p_{other} = 1 - p_A - p_B$ = the estimated proportion of the segment which will not be harvested in either A or B.

Alternatively, estimates may not be made on A and B separately but only on A and B collectively, so that the observation can have the form (p_{A+B}, p_{other}) where

p_{A+B} = the estimated proportion of the segment that will be harvested in either A or B, and

$p_{other} = 1 - p_{A+B}$ = the estimated proportion of the segment that will not be harvested in A or B.

The most recent segment estimates are assumed to reflect any previous observations made on that sample segment during the current growing year. If a sample segment's current observation is of the form (p_{A+B}, p_{other}), then the sample segment is said to be partially classified or partially identified. If its observation is of the form (p_A, p_B, p_{other}), then it will be called completely identified.

The proportion of the region harvested in crop A will be denoted P_A with P_B similarly defined. The objective is to estimate P_A and P_B using the observations on the sample segments. This estimation may have to be made at more than one time during the growing year. Of course, if there are no completely identified sample segments, only the sum $P_A + P_B$ can be estimated on the basis of the sample segments.

Four alternative estimators of the region's at-harvest crop acreage proportions are derived in Technical Report 21

(1) maximum likelihood estimators,

(2) least squares estimators,

(3) weighted least squares estimators, and

(4) a combination of a least squares estimator of the relative proportion of crop A out of crops A and B together and a maximum likelihood estimator of the at-harvest combined acreage proportion of crops A and B together.

The true test of an estimator's value is its performance on real data. Hence a Monte Carlo study of the performance of the four estimation procedures was carried out based upon two real sets of CAMS data.

There were several possible ways to measure the sample behavior of the estimators. For each estimator and each of $\hat{P}_A$, $\hat{P}_B$, and $1 - \hat{P}_A - \hat{P}_B$ the following measures were calculated for each data set:

(i) average absolute error = the average over 1000 simulations of $|\hat{P} - P_{region}|$ where P_{region} represents the actual crop proportion in the particular simulated region.

(ii) average squared error = the average over 1000 simulations of $(\hat{P} - P_{region})^2$

(iii) bias of average estimate = the difference between the average P in 1000 simulations and P_{set} where P_{set} is the actual crop proportion in the entire set of segments, and

(iv) sample variance of the estimator.

Some information is, of course, lost when some segments are only partially identified. To assess this loss, the maximum likelihood estimators were also calculated using the complete identification for all sampled segments. Since these estimators utilize complete information for the entire sample of n segments instead of complete information on only some of the n segments and partial information on the remainder, these latter estimators perform better.

To show that the inclusion of the partially identified segments into the estimation procedure is better than simply ignoring them, the maximum likelihood estimators, least squares estimators, and weighted least squares estimators were also calculated using only the subset of the n sample segments corresponding to the completely identified segments.

In the Monte Carlo Study the CAMS estimates of the segment's crop acreage

proportions were simulated as if they contained no errors. In order to ascertain the impact of any such errors, the Monte Carlo study was repeated with a normal deviate added to each of the segment's crop acreage proportion estimates.

On the basis of the limited Monte Carlo study and the small follow-up investigation the following conclusions were reached:

1) As long as there are some completely identified sample segments, it is reasonable to estimate the individual crop proportions in the region.
2) It is prudent to avoid having a large percentage (say 80%) of only partially identified sample segments.
3) It is much better to incorporate the partially identified sample segments into the estimators than it is to disregard the partially sample segments.
4) When there are either no errors or only very small errors in the estimates of the segment's crop acreage proportions, the maximum likelihood estimators seem to be the best estimators, but they are not greatly superior to weighted least squares estimators or the use of a least squares ratio estimator.
5) When there are fairly substantial errors in the estimates of the segment's crop acreage proportions, the combination of the least squares ratio estimator with the maximum likelihood estimator of the combined crop proportion is the superior estimator.

The overall optimality of using the combination of the least squares ratio estimator and the maximum likelihood estimator of the combined crop proportion suggests some definite possibilities for further research - which we hope to pursue. In particular, by simply treating the combined crops as a single crop,

the combined crop proportion can be estimated when there are more than one year's data by utilizing the current methodology derived for the multiyear model (2) described in section 1. Then this multiyear based combined crop proportion can be subdivided into individual crop proportions using, for example, the least squares ratio estimator based on the current year.

4. Remarks

The productivity of this research period has been aided by the support and cooperation of many NASA, Lockheed, and ERIM personnel. We look forward to our future joint research.

Technical Reports

(18) Gbur, E. E. and R. L. Sielken, Jr. December 1980. Optimal Rotation Designs for Multiyear Estimation, I. Unweighted Estimation.

(19) Gbur, E. E. and R. L. Sielken, Jr. December 1980. Optimal Rotation Designs for Multiyear Estimation, II. Weighted Estimation.

(20) Dahm, P. F. and Robert L. Sielken, Jr. March 1981. Multiyear Estimation of the At-Harvest Crop Acreage Proportion: Methodology and Implementation.

(21) Sielken, Robert L. Jr., December 1981. Incorporating Partially Identified Sample Segments Into Acreage Estimation Procedures: Estimates Using Only Observations From the Current Year.

(22) Gbur, E. E. and R. L. Sielken, Jr. December 1981. Missing Observations in Multiyear Rotation Sampling Designs.